aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
7(2) 109–117

DOI: 10.1177/1745691611432343
http://pps.sagepub.com
SAGE

Revisiting Truth or Triviality: The External Validity of Research in the Psychological Laboratory

Gregory Mitchell
University of Virginia

Abstract

Anderson, Lindsay, and Bushman (1999) compared effect sizes from laboratory and field studies of 38 research topics compiled in 21 meta-analyses and concluded that psychological laboratories produced externally valid results. A replication and extension of Anderson et al. (1999) using 217 lab-field comparisons from 82 meta-analyses found that the external validity of laboratory research differed considerably by psychological subfield, research topic, and effect size. Laboratory results from industrial–organizational psychology most reliably predicted field results, effects found in social psychology laboratories most frequently changed signs in the field (from positive to negative or vice versa), and large laboratory effects were more reliably replicated in the field than medium and small laboratory effects.

Keywords

external validity, generalizability, meta-analysis, effect size

A widely held assumption within the social sciences is that the rigor of experimental research is purchased at the price of generalizability of results (Black, 1955; Locke, 1986; Wilson, Aronson, & Carlsmith, 2010). This trade-off plays out most directly in those fields that use laboratory experiments to study how humans navigate complex social environments, such as in social and industrial–organizational (I-O) psychology. In these fields, highly controlled experiments produce internally valid findings with suspect external validity (e.g., Flowe, Finklea, & Ebbesen, 2009; Greenwood, 2004; Harré & Secord, 1972).

Researchers typically respond to external validity suspicions in one of three ways: by arguing that findings from even highly artificial laboratory studies advance theories that explain behavior outside the laboratory (e.g., Mook, 1983; Wilson et al., 2010), by conducting field studies that demonstrate that causal relations observed in the laboratory hold in the field (e.g., Behrman & Davey, 2001), or by conducting a meta-analysis of laboratory and field studies to assess the impact of research setting on results within a particular area of research (e.g., Avolio, Reichard, Hannah, Walumbwa, & Chan, 2009). Anderson, Lindsay, and Bushman (1999) offered a novel and broad response to the external validity question by comparing 38 pairs of effect sizes from laboratory and field studies of various psychological phenomena as compiled in 21 meta-analyses (i.e., each meta-analysis compared the mean effect size found in the laboratory to that found in the field for the particular phenomenon under investigation).[1] Anderson and colleagues found a high correlation between these meta-analyzed laboratory and field effects ($r = .73$), leading them to conclude that "the psychological laboratory is doing quite well in terms of external validity; it has been discovering truth, not triviality: (Anderson et al., 1999, p. 8).

Anderson et al. (1999) has been widely cited (as of this writing, 150 times in PsycINFO), often for the proposition that psychological laboratory research in general possesses external validity and, thus, the new laboratory finding being reported is likely to generalize (e.g., Ellis, Humphrey, Conlon, & Tinsley, 2006; von Wittich & Antonakis, 2011; West, Patera, & Carsten, 2009). This proposition, and its use to allay external validity concerns about new laboratory findings, assumes the external validity of Anderson and colleagues' conclusion about the external validity of laboratory studies.

However, Anderson and colleagues' conclusion was based on a fairly small number of paired effect sizes that show considerable variation despite the strong overall correlation between laboratory and field results. For instance, their six comparisons of laboratory and field effect sizes from

Corresponding Author:
Gregory Mitchell, School of Law, University of Virginia, Charlottesville, VA 22903.
E-mail: greg_mitchell@virginia.edu

meta-analyses of gender differences in behavior reached inconsistent results ($r = -.03$). Furthermore, their correlational result indicated the direction and magnitude of the relationship, but not the magnitude of differences in effect sizes between the laboratory and the field (i.e., the rank ordering of effects could be quite consistent despite large differences in effect size between the lab and field). Because the small sample examined by Anderson and his colleagues limited the analyses that could be performed and the conclusions that could be drawn from their study, a replication and extension of Anderson et al. (1999) was undertaken to examine the external validity of psychological laboratory research after 10 years using a larger database of effect sizes covering a wider range of psychological phenomena. This larger data set permitted a more detailed examination of external validity by psychological subfield and area of research.[2]

The goal of my study, therefore, was to replicate Anderson et al.'s (1999) study using a larger data set to determine whether their broad positive conclusion about the external validity of laboratory research remains defensible or whether there are identifiable patterns of external validity variation. This study, like Anderson and colleagues' study, is focused on whether laboratory and field results agree and thus employs a coarse distinction between research settings—comparing results obtained under laboratory conditions to those found in the field or under more mundanely realistic conditions. To the extent that variation between the laboratory and field is observed, a more detailed inquiry is called for because many different design variables could account for the variation: differences in participant characteristics between lab and field studies and across cultures (Henrich, Heine, & Norenzayan, 2010; Henry, 2009); differences in guiding design principles such as the use of "mundane realism" versus "psychological realism" (Aronson, Wilson, & Akert, 1994, p. 58) versus representative sampling of stimuli to develop participant tasks, environments, and measures (Dhami, Hertwig, & Hoffrage, 2004); or differences in the timing of the research that may be related to larger societal or historical changes (Cook, 2001). Also, there may be fundamental differences in the generalizability of the processes or phenomena studied across psychological subfields: Some phenomena at some levels of analysis may not vary with the characteristics of the individual and situation, some phenomena may be unique to particular laboratory designs using particular types of participants (i.e., some phenomena may be created in the laboratory rather than be brought into the laboratory for study), and some phenomena may generalize across a narrow range of persons and situations.

In short, examining the consistency of meta-analytic estimates of effects across research settings provides a good first test of the generalizability of laboratory results, but the limits of this approach must be acknowledged. The inferences to be drawn from positive results are limited by the diversity of the participant and situation samples found in the synthesized studies, and negative results call for deeper inquiry into the causes of external invalidity. The meta-analytic data examined here cover a wide range of psychological topics, research settings, and participants. Therefore, if results based on this data set approximate those found by Anderson et al. (1999), then we should have greater confidence in their conclusion that psychological laboratories reveal truths rather than trivialities. If results based on this larger data set differ, then the task will be to understand why some laboratory results generalize while others do not.

Meta-Analytic Data on Effects Studied in the Laboratory and the Field

An effort was made to identify all meta-analyses that synthesized research on some aspect of human psychology conducted in a laboratory setting and in an alternative research setting (see the Appendix for details on the literature search). In keeping with the approach taken by Anderson et al. (1999), comparisons were not limited strictly to laboratory versus field research on the same topic but also included comparisons of results found under less and more mundanely realistic conditions (e.g., the use of experimentally created versus real groups in the study of group behavior and the use of hypothetical versus real transgressions in the study of forgiveness). A review of over 1,100 papers located in the literature search identified 82 meta-analyses reporting effect sizes for at least two research settings, for a total of 217 comparisons of results found under laboratory, or less realistic, conditions to results found under field, or more realistic, conditions (including two dissertations that contributed six lab–field comparisons).[3] The full data set is provided in an online supplement.

Most meta-analyses reported effect sizes in terms of *r*. When an effect size was reported in a unit other than *r*, the effect size was converted to *r* using standard conversion formulas (Cohen, 1988; Rosenthal, 1994). When both weighted and unweighted effect sizes were reported, the weighted effect sizes were used in the analyses reported here.

Four of the meta-analyses compared two types of laboratory studies with one or more types of field studies, and 17 of the meta-analyses compared two or more types of field studies with a single type of laboratory study (see online supplement for details). The results discussed below focus on the comparison of laboratory effects with true field studies or with conditions that differ most from the laboratory conditions because these research settings possess the least "proximal similarity" (Cook, 1990) to the laboratory and thus are likely to raise the greatest generalizability concerns (e.g., McKay & Schare's, 1999, comparison of results found in a traditional laboratory to those found in the field serves as the focal comparison, rather than their comparison of a traditional lab to a "bar lab").[4]

In order to examine possible variation in generalizability across research domains, I classified the meta-analytic data in a number of ways: (a) by PsycINFO group codes that are used to classify studies by primary subject matter (for more information on this classification system, see http://www.apa.org/pubs/

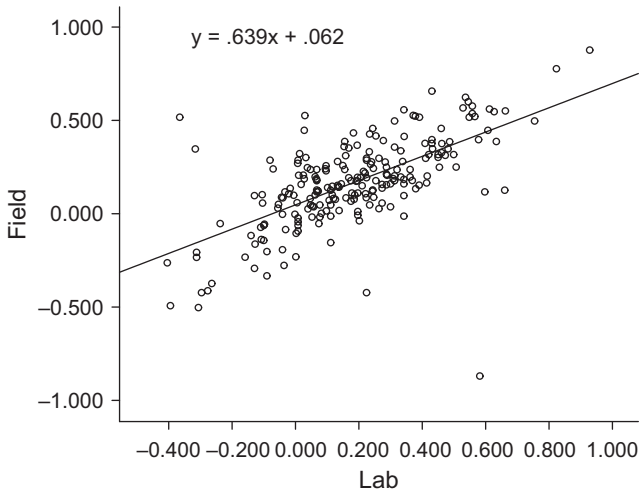

Fig. 1. Scatter plot of paired lab and field effects across all meta-analyses.

databases/training/class-codes.aspx), (b) by psychological subfield as classified by the present author before knowing the PsycINFO classifications of the meta-analyses, (c) by psychological subfield of meta-analysis first author as determined by the affiliation disclosed in the meta-analysis or from information available on the Web if the first author's subfield affiliation was not apparent from the meta-analysis, and (d) by research topics according to PsycINFO subgroup codes and classification by the present author. Results using the PsycINFO classifications are emphasized because those classifications were made by independent coders, show consistency over time, and cover more of the data than some alternative classifications.[5]

Consistency and Variation in Effects in the Laboratory and Field

Aggregate results

A plot of the data reveals considerable correspondence in paired laboratory and field effects (see Fig. 1). When one potential outlier is removed, the overall correlation between lab and field effects in this expanded sample approximates that found in Anderson et al.'s (1999) sample: $r = .71$ versus $r = .73$ reported by Anderson and colleagues (see Table 1 for the full correlation matrix).[6]

As a measure of the reliability of the direction of effects found in the laboratory, the number of times in which a laboratory effect changed its sign in the field (from positive to negative or vice versa) was counted: overall, 30 of 215 laboratory effects changed signs (14%).[7] Thus, a nontrivial number of effects observed in the laboratory produced opposite effects in the field. With respect to the relative magnitude of effects, the mean difference between laboratory and field effects was only .01, but this difference had a standard deviation of .18 on a scale in which the average laboratory and field effects were both $r = .17$.

Results by subfield

It is possible that the dispersion seen in Figure 1 is random across research topics and domains, or it may be that the aggregate results mask systematic differences in lab–field correspondence. To examine possible differences in lab–field correspondence across traditional divisions of psychological inquiry, the paired effects were divided by two alternative subfield classifications: first by the subfield that PsycINFO classified each meta-analysis into, and second by the subfield that I classified each lab–field comparison into (see Table 2). Subfield assignments and results converged under the two approaches to classification, indicating that there was meaning and consistency to the partitioning of the research by psychological subfield.

The two subfields with the greatest number of paired effects, I-O psychology and social psychology, differed considerably in the degree of correspondence between the lab and the field. Laboratory and field effects from I-O psychology correlate very highly ($r = .89$, $n = 72$, 95% CI [.83, .93]), whereas laboratory and field effects from social psychology show a lower correlation ($r = .53$, $n = 80$, 95% CI [.35, .67]).[8] A similar result holds if we partition effects by the subfield affiliation of the first author of each meta-analysis: The

Table 1. Correlation of Lab-Field Effects

	Lab	Lab2	Field	Field2	Field3
Lab 2 (n = 216)	.99 [.99, .99]	—			
Field (n = 216)	.71 [.64, .77]	.70 [.63, .76]	—		
Field 2 (n = 42)	.68 [.48, .82]	.69 [.49, .82]	.57 [.32, .74]	—	
Field 3 (n = 21)	.49 [.07, .76]	.49 [.07, .76]	.63 [.27, .83]	.43 [.00, .73]	—

Note: "Lab" represents collection of primary lab results; "Lab2" substitutes second lab result for primary lab result from four meta-analyses that examined two types of lab studies. "Field" represents collection of primary field results; "Field2" and "Field3" represent field studies from meta-analyses examining two or three different types of field studies. Sample sizes reflect number of paired effect sizes. Brackets present 95% confidence intervals. Results exclude the possible outlier paired-effects from Mullen et al. (1991).

Table 2. Correlation of Lab-Field Effects by Subfield Classifications

PsycINFO classification (*n*)	*r*	*r*	Author's classification (*n*)
Social (80)	.53	.60	Social (79)
I-O (72)	.89	.82	I-O (98)
Personality (22)	.83	.84	Clinical (19)
Consumer (7)	.59	.59	Marketing (7)
Education (7)	.71	.87	Education (5)
Developmental (3)	−.82	−.88	Developmental (6)
Psychometrics/Statistics/Methods (19)	.61		
Human Experimental (5)	.61		

Note: Sample sizes reflect number of paired effect sizes. The PsycINFO classification excludes one pair of effects classified as "Environmental Psychology," and the author classification excludes two pairs of effects classified as "Health Psychology." Results exclude possible outlier effects from Mullen et al. (1991).

lab–field correlation from meta-analyses conducted by I-O authors is .82 (n = 107, 95% CI [.75, .87]), whereas the lab–field correlation from meta-analyses conducted by social psychology authors is .53 (n = 76, 95% CI [.35, .67]).[9]

A plot of paired lab and field effects for I-O psychology and social psychology illustrates the greater convergence of lab and field results within I-O psychology: The slope of the fitted line is steeper for I-O psychology, with I-O lab effects thus being better predictors of field effects (see Fig. 2).[10] Also, the paired effects from I-O psychology differed less in their magnitude, as the distribution around zero difference is steeper for I-O psychology than for social psychology ($Kurtosis_{I\text{-}O}$ = 2.318 vs. $Kurtosis_{Social}$ = −.03). For comparison purposes, a boxplot of the differences in effect size between the laboratory and field across all subfields is provided in Figure 3.

Furthermore, most of the 30 laboratory effects that changed signs in the field came from social psychology. Twenty-one of 80 (26.3%) laboratory effects from social psychology changed signs between research settings, but only 2 of 71 (2.8%) laboratory effects from I-O psychology changed signs; as an additional reference point, only 1 of 22 (.05%) laboratory effects from personality psychology changed signs, $\chi^2(2) = 19.12$, $p < .001$.[11]

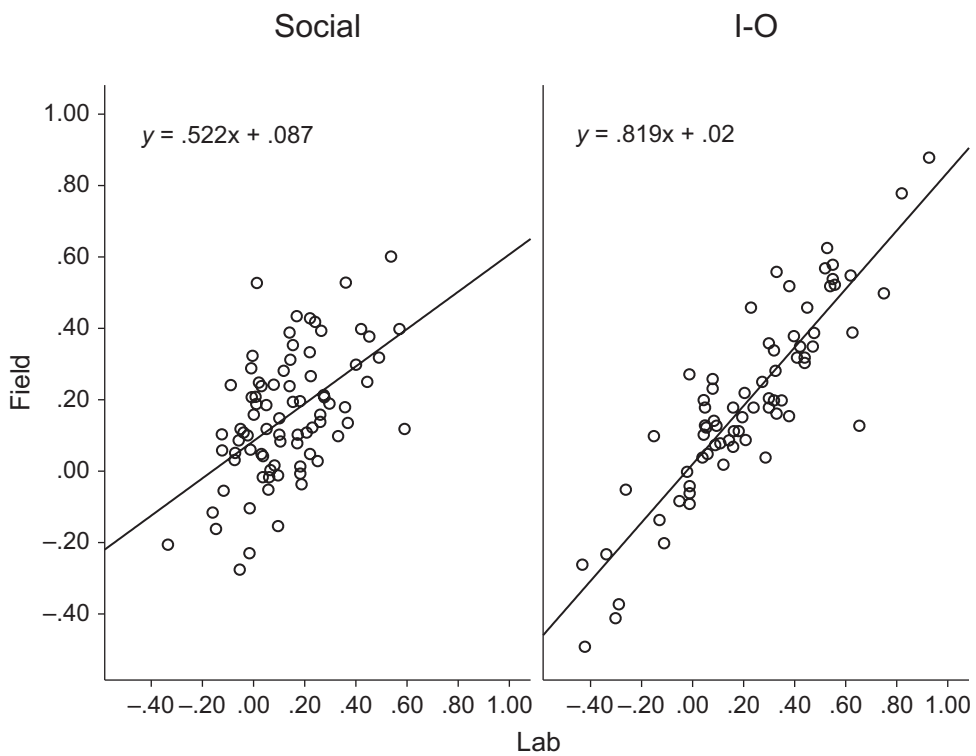

Fig. 2. Scatter plot of paired lab and field effects from social and I-O psychology.

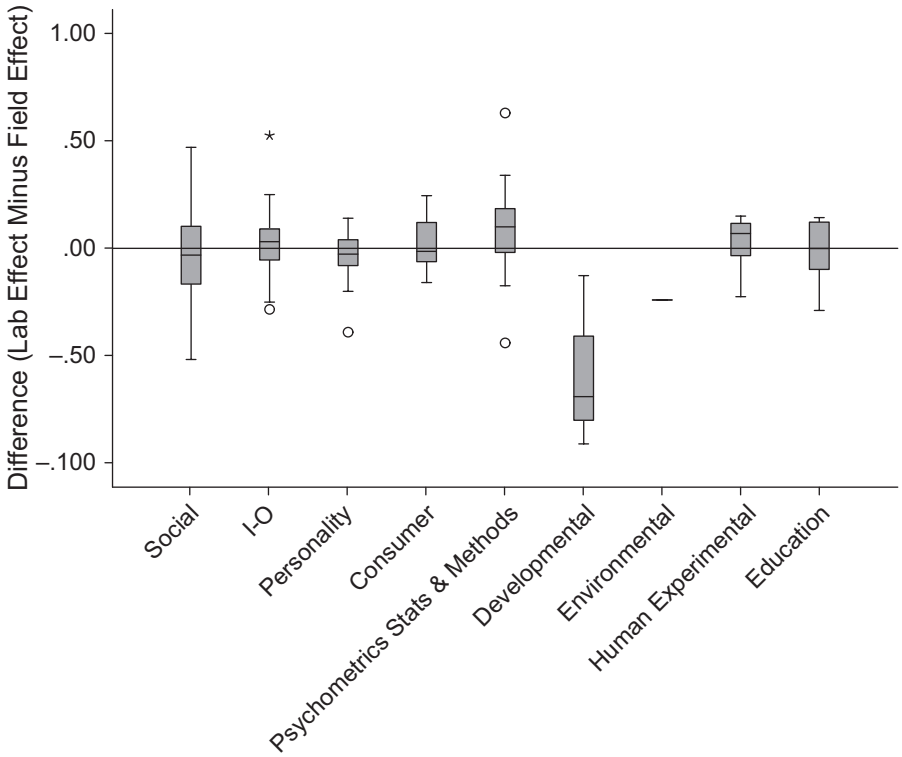

Fig. 3. Boxplot of differences between lab and field effect sizes by subfield.

Results by effect size

A partial explanation for the relatively weaker external validity of social psychology laboratory results appears to be a disproportionate focus on small effect sizes. Using Cohen's rule of thumb to categorize laboratory effect sizes, meta-analyses within I-O psychology examined 29 small, 22 medium, and 21 large laboratory effects, and meta-analyses within social psychology examined 53 small, 20 medium, and 8 large laboratory effects.[12] Small laboratory effects studied by social psychologists varied more in the field than medium effects from social psychology labs: $r_{\text{small effects}} = .30$ ($n = 53$, 95% CI [.03, .53]) vs. $r_{\text{medium effects}} = .57$ ($n = 20$, 95% CI [.17, .81]).[13] Small laboratory effects from I-O psychology likewise varied more in the field than larger effects: $r_{\text{small effects}} = .53$ ($n = 29$, 95% CI [.20, .75]) vs. $r_{\text{medium effects}} = .84$ ($n = 22$, 95% CI [.65, .93]) vs. $r_{\text{large effects}} = .90$ ($n = 21$, 95% CI [.77, .96]). This trend held across all studies, $r_{\text{small effects}} = .47$ ($n = 112$, 95% CI [.31, .60]) vs. $r_{\text{medium effects}} = .56$ ($n = 66$, 95% CI [.37, .71]) vs. $r_{\text{large effects}} = .83$ ($n = 38$, 95% CI [.70, .91]), and small laboratory effects more frequently changed signs in the field than medium and large effects (22.7% vs. 6.1% vs. 2.6%, respectively).

Results by research topic

Lab–field correlations for specific areas of research (e.g., aggression studies, leadership studies) with at least nine meta-analytic comparisons of laboratory and field effects were examined. These results should be interpreted cautiously because they are more sensitive to extreme values given the smaller number of comparisons, but these results do converge with the subfield results because topics of primary interest to I-O psychologists showed the highest correlations and topics of primary interest to social psychologists showed greater variation (see Table 3).

However, these results also illustrate the hazard of assuming that aggregate correlations of lab–field effects are representative of the external validity of all laboratory research within a subfield. There were large differences in the relative magnitude of laboratory and field results across research topics (see the standard deviations in mean effect size differences in Table 3) and in the magnitude of the correlations. For instance, although results from I-O laboratories tended to be good predictors of field results, I-O laboratory studies of performance evaluations were less predictive than I-O laboratory studies of other topics, and leadership studies within I-O psychology were less predictive than leadership studies within social psychology ($r = .63$ for 10 paired laboratory and field effects from leadership meta-analyses conducted by I-O-affiliated authors vs. $r = .93$ for 7 paired effects from leadership meta-analyses conducted by social-affiliated authors). Laboratory studies of gender differences fared particularly poorly compared with other types of social psychological research, which may be due to the small effect sizes found in these studies.[14]

Table 3. Correlation of Lab-Field Effects and Standard Deviations of Effect Size Differences by Research Topic Classifications

Classification (*n*)	*r*	*SD*
PsycINFO classification		
Group Processes & Interpersonal Processes (33)	.58	.18
Social Perception & Cognition (9)	.53	.17
Personality Traits & Processes (20)	.83	.13
Behavior Disorders & Antisocial Behavior [aggression studies] (14)	.68	.14
Personnel Management & Selection & Training (14)	.92	.12
Personnel Evaluation & Job Performance (21)	.74	.16
Organizational Behavior (18)	.97	.09
Author classification		
Aggression-focused comparisons (17)	.63	.13
Gender-focused comparisons (22)	.28	.13
Group-focused comparisons (43)	.63	.19
Leader-focused comparisons (18)	.69	.21

Note: Sample sizes reflect number of paired effect sizes. Results exclude possible outlier effects from Mullen et al. (1991).

Discussion

This expanded comparison of laboratory and field effects replicated Anderson and colleagues' (1999) basic result, but it also raises questions about treating the external validity of psychological laboratory research as an undifferentiated whole: In the aggregate, laboratory and field effect sizes tended to covary (r = .71 vs. Anderson et al.'s r =.73, if we exclude a potential outlier from social psychology), but this result depended on the extremely high correlation of laboratory and field effects from I-O psychology. If we exclude I-O effects, the aggregate correlation drops considerably (to r = .55).

External validity differed across psychological subfields and across research topics within each subfield, and all subfields showed considerable variation in the relative size of effects found in the laboratory versus the field. External validity also differed by effect size: Small laboratory effects were less likely to replicate in the field than larger effects. This latter result empirically demonstrates the importance of considering effect size when planning a field test, not only to determine sample size but also to determine the sensitivity with which measurements should be made and the type of research design needed to isolate the influence of the variables of interest (Cohen, 1988).

Despite the variations in generalizability observed, it is tempting to invoke Cohen's effect size rule of thumb and conclude that all of psychology is performing well in terms of external validity because all subfields showed large lab–field correlations, but doing so would ignore Cohen's (1988) injunction that "the size of an effect can only be appraised in the context of the substantive issues involved" (p. 534). For an investigator considering whether to pursue a new line of research building on prior work, even small lab–field correlations may be sufficient to proceed. For an organization or government agency considering whether to implement a program based on psychological research, even large lab–field correlations may be insufficient, particularly if the costs of implementation are high relative to the likely benefits. To determine likely benefits, the constancy of effect direction and the relative magnitude of the effect in the lab versus that found in the field should be considered, but aggregate correlations between lab and field effects do not provide this information.

Reliance on a subfield's "external validity effect size" could be particularly misleading for results from social psychology, where more than 20% of the laboratory effects changed signs between research settings. Shadish, Cook, and Campbell (2002) emphasize constancy of causal direction over constancy of effect size in their discussion of external validity on grounds that constancy of relations among variables is more important to theory development and the success of applications. The number of sign reversals observed across domains should be cause for concern among those seeking to extend any psychological result to a new setting before any cross-validation work has occurred.

Whether these sign reversals should be cause for concern in any particular case depends on the goals of the research. Mook (1983) correctly noted that some studies require external invalidity to test a prediction or determine what is possible. In such studies, what matters is whether the study helps advance a theory, not whether a specific finding will generalize. But Mook (1983) also noted that, "[u]ltimately, what makes research findings of interest is that they help us understand everyday life" (p. 386). Psychologists often examine minimal, manageable interventions to open a window on psychological processes and causal relations among variables (Prentice & Miller, 1992), and that approach is justifiable if it ultimately produces theories that explain and predict behavior outside the laboratory. Small effects found in the lab can be important, and large effects found in the lab can be unimportant (Cortina & Landis, 2009); whichever is the case must eventually be established in the field.

Conclusion

My results qualify the conclusion reached by Anderson et al. (1999): Many psychological results found in the laboratory can be replicated in the field, but the effects often differ greatly in their size and less often (though still with disappointing frequency) differ in their directions. The pattern of results suggests that there are systematic differences in the reliability of laboratory results across subfields, research topics, and effect sizes, but the reliability of these patterns depends on the representativeness of the laboratory studies synthesized in the meta-analyses that provided the data for this study.

Also, it is possible that alternative divisions of the data would yield different patterns. The data divisions that were

chosen reflect two ideas: (a) different subfields develop and teach unique research design customs and norms (see, e.g., Rozin, 2001), and (b) different research topics require different compromises to enable their study in the laboratory (e.g., prejudice and stereotyping research in the laboratory must often use simulated work situations, whereas research into the accuracy of impressions based on thin slices of behavior may be well-suited for laboratory study;[15] Secord, 1982). Determining the mix of factors responsible for the observed variations in external validity will require further research.

A good starting place for such further inquiry is I-O psychology. Results from I-O labs varied in their generalizability, but the high degree of convergence in I-O effects across research settings indicates that something about this subfield's practices or research topics tends to produce externally valid laboratory research. It may be that I-O psychologist's traditional skepticism of laboratory studies (Stone-Romero, 2002) is adaptive: In a culture that trusts well-done laboratory studies, internal validity challenges will likely command the researcher's (and journal editor's) attention, whereas in a culture that distrusts even well-done laboratory studies, external validity challenges may grab much more of the researcher's (and editor's) attention.[16] It may be that the topics I-O psychologists study are more amenable to laboratory study than those studied by social psychologists, but that seems unlikely given the focus in both subfields on behavior in complex social settings. It may be that I-O psychologists, as primarily applied researchers, benefit from the trial and error of basic researchers in other subfields and are able to devote their attention to robust results. If the explanations all reduce down to the applied focus of I-O psychology, then the external and internal validity of research within the basic research subfields could benefit from greater attention to applications, for replication in the field reduces the chances that relations observed in the laboratory were spurious (Anderson et al., 1999).

Anderson et al. (1999) presented a positive message about the generalizability of psychological laboratory research, but the message here is mixed. We should recognize those domains of research that produce externally valid research, and we should learn from those domains to improve the generalizability of laboratory research in other domains. Applied lessons are often drawn from laboratory research before any cross-validation work has occurred, yet many small effects from the laboratory will turn out to be unreliable, and a surprising number of laboratory findings may turn out to be affirmatively misleading about the nature of relations among variables outside the laboratory.

Appendix

Literature Search

Several exhaustive searches were employed in an effort to locate all meta-analyses of psychological studies in which mean effect sizes in the laboratory and field were computed. First, the EBSCO social science database (which included all psychology journals indexed in the PsycINFO database as well as business, communications, education, health, political science, and sociology journals) and the SAGE psychology database were searched for items with abstracts containing one or more terms from each of the following three sets of terms: (a) *meta-analysis, meta analysis*, *research synthesis*, *systematic review, systematic analysis*, *integrative review*, or *quantitative review*; (b) *lab, laboratory, artificial, experiment, simulation,* or *simulated*; and (c) *field, quasi-experiment, quasi-experimental, real, realistic, real world,* or *naturalistic*. This search was repeated in the PsycINFO database but with the terms allowed to appear in any search field. Another PsycINFO search was conducted for any term from the first list of terms above in the keywords or methodology field and the term *research setting* in any field. These searches produced over 1,100 hits, and the abstracts of all hits were reviewed to eliminate obviously inapplicable materials (e.g., articles focused on research methodology that did not report meta-analytic findings and single studies making reference to meta-analyses of laboratory and field studies) before the texts of hits were examined in detail.

To ensure that the search terms employed in the searches described above did not exclude relevant articles, an additional search was performed in the following journals for any articles containing the term *meta-analysis*: *Academy of Management Journal, Academy of Management Review, American Psychologist, Journal of Experimental Social Psychology, Personnel Psychology*, *Psychological Bulletin, Journal of Applied Psychology, Journal of Social and Personality Psychology, Personality and Social Psychology Bulletin*, and any additional journal within the EBSCO database with *applied, cognition, cognitive psychology,* or *decision* in its publication name.[17] Finally, the reference sections of Richard, Bond, and Stokes-Zoota (2003) and Dieckmann, Malle, and Bodner (2009) and the chapters in Locke (1986) were reviewed for candidates for possible inclusion.

The online supplement, which is provided as a downloadable spreadsheet at http://pps.sagepub.com/supplemental-data lists the meta-analyses included; the research question(s) addressed for each lab–field comparison; and the meta-analytic results for each research setting that was compared, including the number of effects and sample size included in each meta-analytic comparison where this information was reported and the mean effect size associated with each research setting. The supplement also indicates the subfield of psychology into which each meta-analysis was classified by PsycINFO, independently by the present author, and by psychological subfield of the meta-analysis's first author.

Acknowledgments

Hart Blanton, John Monahan, and Fred Oswald provided helpful comments.

Declaration of Conflicting Interests

The author declared that he had no conflicts of interest with respect to his authorship or the publication of this article.

Notes

1. It is more accurate to say that Anderson, Lindsay, and Bushman (1999) primarily compared effects in the lab with those in the field; they did not strictly limit their comparisons to lab versus field studies but also compared findings for real versus artificial groups and for real versus hypothetical events.
2. Proctor and Capaldi (2001) called for an extension of Anderson et al. (1999) to include more research domains, but no such extension has previously been reported.
3. There are a few meta-analyses that examined effects under different research settings, but they could not be included because they did not report effect size information for each of the settings (e.g., Frattaroli, 2006).
4. The results also include those meta-analyses that had some overlap in coverage (these overlapping meta-analyses are identified in the notes to the online supplement). None of the results differ greatly if the earlier of the overlapping meta-analyses are excluded (e.g., aggregate lab-field r =.64 with overlapping studies included and excluded).
5. For instance, a journal-based approach to classifying research by subfields (e.g., comparing traditional social to I–O journals) leads to a loss of data because several meta-analyses from different subfields were published in *Psychological Bulletin*. Nevertheless, every alternative classification of the effects examined produced results similar to those reported here, including classification of the effects by journal subfield.
6. One set of paired effects from Mullen et al. (1991) comparing the effect of interpersonal distance on permeability of group boundaries in imaginary and real groups showed an extreme disparity between lab and field results (see the lower right quadrant of Fig. 1). Accordingly, the results reported in the text do not include this pair of effects. With Mullen et al. included in the analysis, the overall r = .64.
7. This count excluded two comparisons (one from social and one from I–O) in which one of the paired effect sizes equaled zero.
8. Mullen et al. (1991) fell within the domain of social psychology; with Mullen et al. included in this analysis, the correlation for social psychology drops to r = .29 (n = 81, 95% CI [.08, .48]).
9. The first author of Mullen et al. (1991) was a social psychologist; with Mullen et al. included in this analysis, the correlation for social psychology drops to r = .27 (n = 77, 95% CI [.05, .47]).
10. When Mullen et al. (1991) is included in the social psychology effects, $y = .325x + .098$.
11. Four of 19 paired effects within PsycINFO's "Psychometrics & Statistics & Methodology" classification changed signs (21%), but meta-analyses in this method-focused classification implicated subject matter from other subfields (the four sign reversals within this classification involved the impact of test expectancies on multiple-choice tests, the relation of two different aspects of leader styles to work performance, and the impact of question wording on causal attributions for success and failure). Using my subfield classifications, which distributed these 19 studies into other subject matter subfields, 18 of 80 (23%) social psychology comparisons, 8 of 96 (8%) I–O psychology comparisons, and 1 of 19 (5%) clinical psychology comparisons produced sign changes, $\chi^2(2) = 8.64$, p = .013.
12. Lab effect sizes were categorized based on Cohen's (1988) rule of thumb for the size of correlation coefficients (small r = .10, medium r = .30, and large r = .50) using the following ranges: small effects are absolute effect sizes of .20 or less, medium effects are absolute effect sizes from .201 to .40, and large effects are absolute effect sizes of .401 or greater.
13. Only eight large laboratory effect sizes were found for social psychology, one of which was the possible outlier; the lab–field correlation based on the remaining seven large effects from social psychology laboratories (r = −.13) is thus susceptible to considerable influence by new results.
14. With gender studies excluded, the lab–field correlation increases slightly for social psychology (from r = .53 to r = .56) and does not change for I-O psychology (r = .89).
15. Suitability for study in the lab does not ensure generalizability; many factors on the design side will also come into play (Dhami, Hertwig, & Hoffrage, 2004; Hammond, Hamm, & Grassia, 1986).
16. Attempts to pre-empt external validity challenges may explain why laboratory studies of aggression by social psychologists performed better in the field than some other areas of social psychological research. Aggression researchers have long faced skepticism about their work's applied implications (Berkowitz & Donnerstein, 1982); indeed, such skepticism seems to have been part of the reason for the study by Anderson et al. (1999).
17. Only post-1998 issues of *Psychological Bulletin*, *Journal of Applied Psychology*, *Journal of Social and Personality Psychology*, and *Personality and Social Psychology Bulletin* were searched to supplement the relevant articles found in pre-1999 issues of these journals by Anderson et al. (1999).

References

Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality? *Current Directions in Psychological Science, 8*, 3–9.

Aronson, E., Wilson, T. D., & Akert, R. M. (1994). *Social psychology: The heart and mind*. New York, NY: Harper Collins.

Avolio, B. J., Reichard, R. J., Hannah, S. T., Walumbwa, F. O., & Chan, A. (2009). A meta-analytic review of leadership impact research: Experimental and quasi-experimental studies. *Leadership Quarterly, 20*, 764–784.

Behrman, B. W., & Davey, S. L. (2001). Eyewitness identification in actual criminal cases: An archival analysis. *Law and Human Behavior, 25*, 475–491.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist, 37*, 245–257.

Black, V. (1955). Laboratory versus field research in psychology and the social sciences. *British Journal for the Philosophy of Science, 5*, 319–330.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook, T. D. (1990). The generalization of causal connections: Multiple theories in search of clear practice. In L. Sechrest, E. Perrin,

& J. Bunker (Eds.), *Research methodology: Strengthening causal interpretations of nonexperimental data* (DHHS Publication No. PHS 90-3454, pp. 9–31). Rockville, MD: U.S. Department of Health and Human Services.

Cook, T. D. (2001). Generalization: Conceptions in the social sciences. In N. J. Smelser, J. Wright, & P. B. Baltes (Eds.), *9 International encyclopedia of the social and behavioral sciences* (pp. 6037–6043). Oxford, UK: Pergamon-Elsevier.

Cortina, J. M., & Landis, R. S. (2009). When small effect sizes tell a big story, and when large effect sizes don't. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends* (pp. 287–308). New York, NY: Routledge.

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130*, 959–988.

Dieckmann, N. F., Malle, B. F., & Bodner, T. E. (2009). An empirical assessment of meta-analytic practice. *Review of General Psychology, 13*, 101–115.

Ellis, A. K. J., Humphrey, S. E., Conlon, D. E., & Tinsley, C. H. (2006). Improving customer reactions to electronic brokered ultimatums: The benefits of prior experience and explanations. *Journal of Applied Social Psychology, 36*, 2293–2324.

Flowe, H. D., Finklea, K. M., & Ebbesen, E. B. (2009). Limitations of expert psychology testimony on eyewitness identification. In B. L. Cutler (Ed.), *Expert testimony on the psychology of eyewitness identification* (pp. 201–221). New York, NY: Oxford University Press.

Frattaroli, J. (2006). Experimental disclosure and its moderators: A meta-analysis. *Psychological Bulletin, 132*, 823–865.

Greenwood, J. D. (2004). What happened to the "social" in social psychology? *Journal for the Theory of Social Behaviour, 34*, 19–34.

Hammond, K. R., Hamm, R. M., & Grassia, J. (1986). Generalizing over conditions by combining the multitrait-multimethod matrix and the representative design of experiments. *Psychological Bulletin, 100*, 257–269.

Harré, R., & Secord, P. F. (1972). *The explanation of social behavior*. Lanham, MD: Rowman & Littlefield.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61–83.

Henry, P. J. (2009). College sophomores in the laboratory redux: Influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry, 19*, 49–71.

Locke, E. A. (Ed.). (1986). *Generalizing from laboratory to field settings*. Lexington, MA: Lexington Books.

McKay, D., & Schare, M. L. (1999). The effects of alcohol and alcohol expectancies on subjective reports and physiological reactivity: A meta-analysis. *Addictive Behaviors, 24*, 633–647.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*, 379–387.

Mullen, B., Copper, C., Cox, P., Fraser, C., Hu, L., Meisler, A., . . . Symons, C. (1991). Boundaries around group interaction: A meta-analytic integration of the effects of group size. *The Journal of Social Psychology, 131*, 271–283.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112*, 160–164.

Proctor, R. W., & Capaldi, E. J. (2001). Empirical evaluation and justification of methodologies in psychological science. *Psychological Bulletin, 127*, 759–772.

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*, 331–363.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York, NY: Russell Sage Foundation.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review, 5*, 2–14.

Secord, P. F. (1982). The behavior identity problem in generalizing from experiments. *American Psychologist, 37*, 1408.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Stone-Romero, E. F. (2002). The relative validity and usefulness of various empirical research designs. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 77–98). Malden, MA: Blackwell.

von Wittich, D., & Antonakis, J. (2011). The KAI cognitive style inventory: Was it personality all along? *Personality and Individual Differences, 50*, 1044–1049.

West, B. J., Patera, J. L., & Carsten, M. K. (2009). Team level positivity: Investigating positive psychological capacities and team level outcomes. *Journal of Organizational Behavior, 30*, 249–267.

Wilson, T. D., Aronson, E., & Carlsmith, K. (2010). The art of laboratory experimentation. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 1, pp. 51–81). Hoboken, NJ: Wiley.